Original Paper

Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study

Amanda L Zaleski[1,2], MSc, PhD; Rachel Berkowsky[3], MSc; Kelly Jean Thomas Craig[1], PhD; Linda S Pescatello[3], MSc, PhD

[1]Clinical Evidence Development, Aetna Medical Affairs, CVS Health Corporation, Hartford, CT, United States

[2]Department of Preventive Cardiology, Hartford Hospital, Hartford, CT, United States

[3]Department of Kinesiology, University of Connecticut, Storrs, CT, United States

Corresponding Author:
Amanda L Zaleski, MSc, PhD
Clinical Evidence Development
Aetna Medical Affairs
CVS Health Corporation
151 Farmington Avenue
Hartford, CT, 06156
United States
Phone: 1 8605385003
Email: zaleskia@aetna.com

Abstract

Background: Regular physical activity is critical for health and disease prevention. Yet, health care providers and patients face barriers to implement evidence-based lifestyle recommendations. The potential to augment care with the increased availability of artificial intelligence (AI) technologies is limitless; however, the suitability of AI-generated exercise recommendations has yet to be explored.

Objective: The purpose of this study was to assess the comprehensiveness, accuracy, and readability of individualized exercise recommendations generated by a novel AI chatbot.

Methods: A coding scheme was developed to score AI-generated exercise recommendations across ten categories informed by gold-standard exercise recommendations, including (1) health condition–specific benefits of exercise, (2) exercise preparticipation health screening, (3) frequency, (4) intensity, (5) time, (6) type, (7) volume, (8) progression, (9) special considerations, and (10) references to the primary literature. The AI chatbot was prompted to provide individualized exercise recommendations for 26 clinical populations using an open-source application programming interface. Two independent reviewers coded AI-generated content for each category and calculated comprehensiveness (%) and factual accuracy (%) on a scale of 0%-100%. Readability was assessed using the Flesch-Kincaid formula. Qualitative analysis identified and categorized themes from AI-generated output.

Results: AI-generated exercise recommendations were 41.2% (107/260) comprehensive and 90.7% (146/161) accurate, with the majority (8/15, 53%) of inaccuracy related to the need for exercise preparticipation medical clearance. Average readability level of AI-generated exercise recommendations was at the college level (mean 13.7, SD 1.7), with an average Flesch reading ease score of 31.1 (SD 7.7). Several recurring themes and observations of AI-generated output included concern for liability and safety, preference for aerobic exercise, and potential bias and direct discrimination against certain age-based populations and individuals with disabilities.

Conclusions: There were notable gaps in the comprehensiveness, accuracy, and readability of AI-generated exercise recommendations. Exercise and health care professionals should be aware of these limitations when using and endorsing AI-based technologies as a tool to support lifestyle change involving exercise.

(JMIR Med Educ 2024;10:e51308) doi: 10.2196/51308

KEYWORDS

exercise prescription; health literacy; large language model; patient education; artificial intelligence; AI; chatbot

Introduction

Regular physical activity is an essential component of a healthy lifestyle with numerous benefits that are widely recognized and indisputable [1,2]. To support overall health, the American College of Sports Medicine (ACSM) and the Department of Health and Human Services recommend healthy adults engage in regular physical activity, including moderate-intensity aerobic exercise for at least 150 minutes per week, vigorous-intensity aerobic exercise for at least 75 minutes per week, or a combination of both, as well as muscle-strengthening activities at least twice per week [1,2]. In addition, evidence-based practice calls for exercise as first-line therapy to prevent, treat, and control multiple chronic conditions and diseases such as hypertension, hypercholesterolemia, and diabetes mellitus [3-7]. As such, the ACSM endorses individualized, evidence-based, exercise recommendations (termed *exercise prescription* [ExRx]) for more than 25 clinical populations [1]. These ExRxs are tailored to favorably augment health-related outcomes of interest for each respective clinical population while addressing additional factors such as clinical contraindications, common medications, and special considerations [1,8]. Despite well-established guidelines, health care providers often struggle to provide sufficient counseling and follow-up on lifestyle recommendations, including exercise, due to various barriers such as time constraints, limited resources, lack of awareness or training, and lack of reimbursement incentives [9-11]. Patients also rely heavily on web-based sources for health-related information [12-14], which often includes misinformation that can negatively impact health outcomes and undermine provider-led efforts to support behavior change [15,16].

Artificial intelligence (AI) has recently emerged as a promising tool to augment health and health care and address these challenges [17]. AI-based technology including machine learning, neural networking, deep learning, and natural language processing enables computers to interact with a corpus of text data to generate human language [18,19]. Large language models (LLMs), such as the generative pretrained transformer (GPT), have the ability to generate human-like language on their own, making them a powerful tool for interacting with users as if they are communicating with another human [18,19]. The surge in popularity of LLMs can largely be attributed to the third iteration of OpenAI's GPT series, ChatGPT [20]. ChatGPT has been recognized as the fastest-growing consumer application in history [20] and is widely regarded as disruptive technology due to its strong potential to enable a wide range of clinical applications as both a provider- and patient-facing tool [21] by generating language that is contextually appropriate, natural sounding, and coherent. Indeed, ChatGPT has demonstrated remarkable capabilities including diagnosis support, streamlining clinical workflows, reducing documentation burden, improving patient education understandability and experience [22-25], and, most recently, passing the United States Medical Licensing Examination [26].

Transformative applications of ChatGPT continue to evolve, but evaluation of its output and suitability in clinical context remains to be explored, in addition to identifying barriers to access and outcomes related to its use. The application of digital technology to support a health behavior change using knowledge-shaping techniques, which is complex and riddled with contextual and individualized components, is challenging [27]. Challenges include ensuring the suitability and usability of the technology confers appropriate educational requisites to understand and apply knowledge in the form of its recommendations. These educational considerations include readability, which can influence the use of AI-generated education for health behavior change [28]. Further, as an extension of readability, low health literacy can limit a patient's ability to understand and use health information effectively, which can reduce the effectiveness of AI-generated educational resources [29,30].

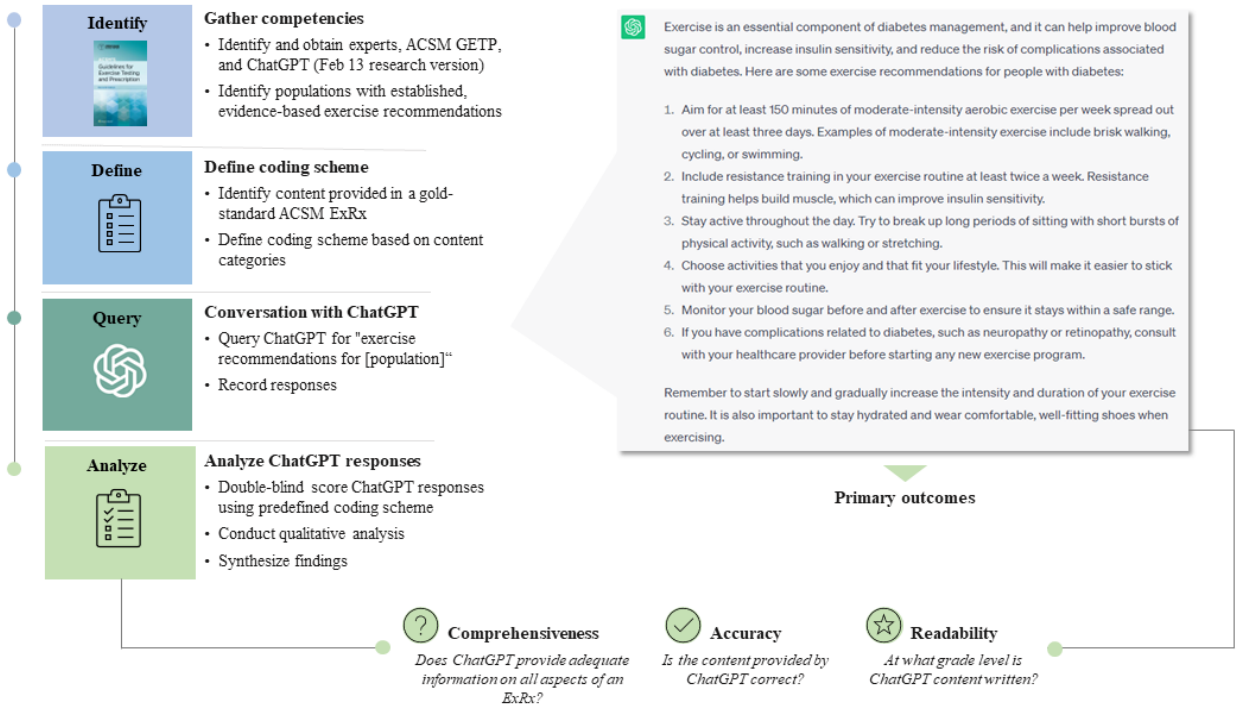
The evaluation of ChatGPT's suitability to provide interactive, personalized, and evidence-based exercise recommendations to support behavior change to improve health has not been conducted to date. As such, the primary aim of this study is to assess the suitability of exercise recommendations generated by ChatGPT, a new AI chatbot, as an adjuvant educational tool for health care providers and patients. Primary outcomes of interest include comprehensiveness, accuracy, and readability of the recommendations generated by ChatGPT, with the goal of determining its potential to deliver personalized exercise recommendations at scale. A secondary aim of this study was to conduct a qualitative analysis to identify potential patterns, consistencies, and gaps in AI-generated exercise recommendations. As this technology is still nascent, the study was exploratory in nature, without an a priori hypothesis.

Methods

High-Level Overview

This study was conducted in March 2023 using the free research preview of a novel AI chatbot (ChatGPT February 13 version) [31]. Figure 1 provides a conceptual overview of the study. Briefly, open-text queries seeking individualized exercise advice were posed to the chatbot interface for all populations (N=26) for which there exist established evidence-based exercise recommendations by the ACSM [1]. Mixed methods were applied to characterize individual and average exercise recommendation content depth, accuracy, and readability. The results were synthesized to highlight potential strengths, weaknesses, opportunities, and risks for researchers, clinicians, and patients likely to interact with the ChatGPT platform for this use case.

Figure 1. Conceptual study overview. ACSM: American College of Sports Medicine; ExRx: exercise prescription; GETP: Guidelines for Exercise Testing and Prescription.



Ethical Considerations

This study was deemed to be exempt by the University of Connecticut Institutional Review Board (E23-0378) as this study solely involved the evaluation of AI-generated output and did not involve interaction or intervention with human subjects.

Selection of the Gold-Standard Reference Source

The ACSM is widely regarded as a leading authority in the field of exercise science and sports medicine, and the organization's guidelines and recommendations are considered the gold standard for health and fitness professionals in the United States and the world [1,8,32]. The *ACSM's Guidelines for Exercise Testing and Prescription* (GETP) serves as its flagship resource manual, continuously updated every 4-5 years since 1975. The most recent edition integrates the latest guidelines from ACSM position stands and other relevant professional organizations' scientific statements, including the 2018 Physical Activity Guidelines for Americans [1]. This latest edition of GETP represents the most current and primary resource for evidence-based exercise recommendations [1]. Given ACSM's authoritative status and the comprehensiveness of its guidelines, GETP was selected as the ground truth benchmark source to guide the study design and systematically evaluate the suitability of AI-generated exercise recommendations.

ChatGPT Prompt Specificity and Structure

Prompt methodology was developed a priori with the overarching goal to observe ChatGPT's unaltered performance in a real-world setting while controlling for factors known to influence output, including prompt structure, evaluation timeframe, model version, and model feedback.

A single researcher (ALZ) posed 26 separate, open-ended prompts as a new chat session to the ChatGPT bot (version 3.5) on the same day in a single session. Each text prompt was framed to the ChatGPT bot in a standardized, neutral, third-person tense format as "exercise recommendations for [population]" to optimize the relevance of AI responses for both health care provider and patient scenarios. Generated ChatGPT bot responses were abstracted from the interface and converted into plain text format using Microsoft Word (version 2208; Microsoft Corp) on the same day. Content was unaltered upon conversion to plain text format (Multimedia Appendix 1). Note that the ChatGPT bot used in this study was not subjected to retraining or correction during these prompt interactions. The rationale for this methodological decision was to enable the natural observation of ChatGPT's raw performance and provide a transparent evaluation of its inherent capabilities [33,34].

AI-Generated Exercise Recommendations

Following this prompt specificity and structure, all clinical populations within the ACSM GETP were evaluated once in a separate prompt (N=26), including healthy adults, children and adolescents, older adults, persons who are pregnant, and individuals with cardiovascular disease (CVD), heart failure, heart transplant, peripheral artery disease, cerebrovascular accident, asthma, chronic obstructive pulmonary disease, diabetes mellitus, dyslipidemia, hypertension, overweight and obesity, arthritis, cancer, fibromyalgia, HIV, kidney disease, multiple sclerosis, osteoporosis, spinal cord injury, Alzheimer disease, intellectual disability, and Parkinson disease.

Conceptual Content Analysis

A list of conceptual categories was generated, refined, and organized into a coding scheme for predefined categories that pertain to the fundamental aspects of an ExRx. These categories relate to an individualized physical activity program based on the FITT principle, which stands for the frequency (*how often?*),

intensity (*how hard?*), time (*how long?*), and type (*what kind?*) of exercise [1,35]. The final coding scheme included ten categories: (1) health condition–specific benefits of exercise, (2) exercise preparticipation health screening, (3) frequency, (4) intensity, (5) time, (6) type, (7) volume, (8) progression, (9) special considerations, and (10) references (ie, citations to primary literature or sources that supported the AI-generated content provided).

AI-generated exercise recommendations were then coded and recorded in Microsoft Excel (version 2208; Microsoft Corp) following a 2-stage coding process by 2 independent coders with advanced degrees in kinesiology (ALZ and RB). In the first stage, AI-generated content was appraised for comprehensiveness. Each exercise recommendation was coded for the presence (1 point) or absence (0 points) of content provided for each of the 10 prespecified categories such that each exercise recommendation had a possible range of 0-10 points. Comprehensiveness was determined by dividing the total number of points (ie, *actual*) by the total number possible (ie, *expected* or 10 points) and multiplying by 100. The resulting score was expressed as a percent, with 100% indicating the highest possible score and fully comprehensive. This formula was applied to all 26 exercise recommendations and averaged to characterize ChatGPT's overall ability to deliver exercise recommendations regarding their comprehensiveness.

In the second stage, all categories with reported content (ie, fully *and* partially comprehensive content) were appraised for accuracy. Accuracy was defined as concordance with the ACSM GETP as the ground truth source [1]. In one instance, content deviated from the ACSM GETP (ie, condition-specific benefits of exercise for individuals with HIV), and accuracy was defined as the degree to which the content was consistent with other widely established facts or clinical literature. Responses were coded by the same independent reviewers (ALZ and RB) and recorded as binary variables: "concordant" or "discordant" following the same process used to determine comprehensiveness. Potential discrepancies in coding were resolved through discussion with a third party and senior expert in the field (LSP). The accuracy score was determined by dividing the number of concordant category counts by the number of categories present (ie, "actual" counts; previously determined when calculating comprehensiveness during the first stage) and multiplying by 100. The resulting score was expressed as a percent, with 100% indicating the highest possible accuracy score or fully concordant.

Readability Metrics

The Flesch-Kincaid formula was used to determine readability, a commonly used tool that evaluates the complexity of text-based educational material. This tool was selected due to its objectivity, as scores are computationally derived rather than paper-and-pencil tools that rely on hand calculations and subjectivity, which introduce risk for human error [36]. The formula is based on the average number of syllables per word and the average number of words per sentence with the resulting score estimating the minimum grade level required to understand the text. For example, a score of 8.0 means that the text can be understood by an average eighth-grader in the United States. Flesch reading ease scores range from 0 to 100, with higher scores indicating easier-to-read text. For example, scores <50 are considered difficult to read, while scores >80 are considered easy to read [36]. To assess readability metrics and word count, a single researcher (RB) used the built-in readability statistics functionality of Microsoft Word (version 2208). The mean (SD) word count and readability metrics (ie, Flesch reading ease and grade level) were calculated using Microsoft Excel (version 2208).

Qualitative Analysis

Qualitative analysis with a thematic mapping approach was used to identify novel patterns, trends, and insights across the AI-generated text output. Thematic mapping, a qualitative research method, involves the identification, analysis, and visualization of recurring themes or topics within a data set. This approach is instrumental in highlighting consistencies or gaps in data, facilitating the generation of insights, and formulating hypotheses for further investigation [37].

Statistical Analyses

Descriptive statistics characterized the distribution of all outcome variables of interest, including comprehensiveness, accuracy, and readability metrics. Interrater reliability was assessed using Cohen κ coefficient [(observed agreement–expected agreement)/(1–expected agreement)]. Qualitative analysis was conducted using a systematic multistep approach. All AI-generated exercise recommendations, comprising the text output, were collected and organized to form the data set for qualitative examination. The analysis was carried out by a single researcher (ALZ) who immersed themselves in the content and initiated the coding process by identifying initial themes or patterns within the recommendations. Subsequently, codes were meticulously refined and organized into broader themes, ensuring consistency and accuracy throughout the process. These identified themes were then visually mapped to represent patterns within the data set. Insights generated from the analysis were discussed collaboratively as a team, facilitating comprehensive understanding and quantification, whenever applicable.

Results

Interrater Reliability

Interrater reliability was assessed for the 2 independent raters who coded a sample of 26 AI-generated exercise recommendations using a set of 10 categories. Cohen κ coefficient was calculated to be 1.0, indicating perfect agreement between coders.

Comprehensiveness of AI-Generated Exercise Recommendations

Table 1 details the presence of educational content across the predefined categories of interest abstracted from AI-generated exercise recommendations for 26 populations. Overall, AI-generated exercise recommendations were 41.2% (107/260) comprehensive when compared against a predefined set of content categories that comprise a gold-standard ExRx [1]. There were no populations or categories that were fully

comprehensive. Comprehensiveness ranged from 0% to 92% with notable gaps in content surrounding the critical components of ExRx: frequency (n=2, 8%), intensity (n=2, 8%), time (n=1, 4%), and volume (n=0, 0%). Partial information was provided across these same categories (ranging from 31% to 58%) with almost all gaps surrounding the provision of FITT for resistance training or flexibility modalities. In addition, only 8% (n=2) of recommendations provided a reference source, both of which (accurately) cited the American Heart Association.

Table 1. Comprehensiveness of artificial intelligence–generated exercise recommendations by content category (N=26).

Content	Exercise recommendations reporting content		
	Fully provided, n (%)	Partial[a], n (%)	Not provided, n (%)
Condition-specific benefits	24 (92)	0 (0)	2 (8)
Preparticipation screening	24 (92)	0 (0)	2 (8)
Frequency	2 (8)	9 (35)	15 (58)
Intensity	2 (8)	15 (58)	9 (35)
Time	1 (4)	10 (38)	15 (58)
Type	14 (54)	12 (46)	0 (0)
Volume	0 (0)	8 (31)	18 (69)
Progression	15 (58)	0 (0)	11 (42)
Special considerations	23 (88)	0 (0)	3 (12)
References	2 (8)	0 (0)	24 (92)

[a]Partial indicates some, but not all, possible content was provided.

Accuracy of AI-Generated Exercise Recommendations

Of the total available content provided to the end user, AI-generated exercise recommendations were 90.7% (146/161) accurate when compared to a gold-standard reference source (ie, ACSM GETP [1]). Among the 9.3% (15/161) of inaccurate recommendations (Table 2), there were 15 counts of discordance with most misinformation counts (n=8, 53%) surrounding the need for preparticipation medical clearance prior to engaging in exercise. The second highest category of discordance was within education related to frequency (n=2, 13%) with "overprescribing" aerobic exercise for Alzheimer disease and fibromyalgia by 2 and 5 days per week, respectively. There was 1 count each of discordance across 5 content categories (ie, condition-specific benefits, intensity, time, type, and progression) and 0 counts of discordance across the remaining content categories, including volume, special considerations, and references.

When comparing populations with discordance, hypertension (n=3) had the greatest number of misinformation counts followed by individuals with fibromyalgia (n=2), healthy adults (n=1), older adults (n=1), and cancer (n=1) and individuals with Alzheimer disease (n=1), arthritis (n=1), dyslipidemia (n=1), HIV (n=1), multiple sclerosis (n=1), osteoporosis (n=1), and overweight and obesity (n=1).

Table 2. Summary of inaccurate content among AI[a]-generated exercise recommendations for all reported content categories (N=161).

Content category (count/reported) and population	AI-generated content	Ground truth
Condition-specific benefits (1/24)		
HIV	"Regular exercise can be beneficial for people living with HIV. It can help improve cardiovascular health, reduce stress and anxiety, boost immune function, and enhance overall well-being."	Research regarding the impact of exercise to improve immune function is inconclusive [38].
Preparticipation screening (8/24)		
Healthy adult	"Remember to always listen to your body and consult with a healthcare professional before starting a new exercise routine."	ACSM[b] preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Older adult	"However, it's important for older adults to choose exercises that are appropriate for their age and fitness level, and to check with their doctor before starting a new exercise program."	ACSM preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Dyslipidemia	"It's important to consult with a healthcare professional before starting a new exercise program ..."	ACSM preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Hypertension	"Consult with your doctor: Before starting any exercise program, it is important to consult with your doctor to make sure it is safe for you to exercise and to get guidance on the most appropriate type and intensity of exercise."	ACSM preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Arthritis	"Be sure to talk to your doctor or a physical therapist before starting any new exercise program ..."	ACSM preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Cancer survivors	"It's important to work with a healthcare provider or a qualified exercise professional to develop a safe and effective exercise program that meets your individual needs and abilities."	ACSM preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Multiple sclerosis	"It is always recommended to consult with a healthcare professional before starting any exercise program."	ACSM preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Osteoporosis	"It's important to talk to your doctor or a qualified exercise professional before starting any new exercise program, especially if you have osteoporosis or other medical conditions."	ACSM preparticipation screening guidelines state that "physically inactive but otherwise healthy asymptomatic persons may begin light- to moderate-intensity exercise without medical clearance and, in the absence of symptoms, progress gradually in intensity as recommended by current ACSM exercise prescription guidelines" [1].
Frequency (2/11)		
Fibromyalgia	"Aim for at least 30 minutes of aerobic exercise most days of the week."	ACSM recommends an initial frequency of 1-2 days per week, gradually progressing to 2-3 days per week [1].

Content category (count/reported) and population	AI-generated content	Ground truth
Alzheimer disease	"Engage in moderate aerobic exercise such as brisk walking, cycling, or swimming for at least 30 minutes a day, five days a week."	ACSM recommends a frequency of 3 days per week [1].
Intensty (1/17)		
Hypertension	"Avoid high-intensity exercises: Avoid high-intensity exercises that can cause sudden increases in blood pressure, such as sprinting or heavy lifting."	ACSM does not contraindicate vigorous-intensity aerobic exercise or heavy lifting assuming adequate progression, absence of underlying disease, and proper breathing technique (ie, avoidance of Valsalva maneuver) [1].
Time (1/11)		
Fibromyalgia	"Start with 1-2 sets of 10-15 repetitions for each exercise and gradually increase the resistance as tolerated."	ACSM recommends gradual progression of 4-5 to 8-12 repetitions and increasing from 1 to 2-4 sets per muscle group [1].
Type (1/26)		
Hypertension	"Aim for at least 30 minutes of moderate-intensity aerobic exercise most days of the week."	New ACSM guidelines reinforce that emphasis is no longer placed on aerobic exercise alone. Aerobic or resistance exercise alone or aerobic and resistance exercise combined (ie, concurrent exercise) is recommended on most, preferably all, days of the week to total 90 to 150 minutes per week or more of multimodal, moderate-intensity exercise [39].
Volume (0/8)		
N/A[c]	N/A	N/A
Progression (1/15)		
Overweight and obesity	"If you're new to exercise, start with low-intensity activities such as walking or swimming, and gradually increase your intensity and duration."	ACSM recommends initial intensity should be moderate, progressing to vigorous for greater health benefits [1].
Special considerations (0/23)		
N/A	N/A	N/A
References (0/2)		
N/A	N/A	N/A

[a]AI: artificial intelligence.

[b]ACSM: American College of Sports Medicine.

[c]N/A: not applicable.

Readability Metrics

Average and individual readability metrics and word count for AI-generated exercise recommendations are provided in Table 3. On average, AI-generated output was 259.3 (SD 49.1) words (range 171-354) and considered "difficult to read" with an average Flesch reading ease of 31.1 (SD 7.7; range 14.5-47.3) and written at a college-level (mean 13.7, SD 1.7; range 10.1-18.0).

Table 3. Readability metrics for artificial intelligence–generated exercise recommendations by population.

Population	Word count	Flesch reading ease	Grade level
Healthy adults	187	14.5	15.2
Children and adolescents	253	29.8	14.1
Pregnancy	267	34.7	13.5
Older adults	276	37.0	12.2
Cardiovascular disease	271	33.6	13.2
Heart failure	235	23.0	16.2
Heart transplant	278	24.9	14.4
Peripheral artery disease	322	32.4	13.4
Cerebrovascular accident	346	22.0	15.1
Asthma	317	41.1	12.0
COPD[a]	247	47.3	10.1
Diabetes	201	36.7	11.8
Dyslipidemia	291	19.6	15.9
Hypertension	247	34.5	13.3
Overweight and obesity	200	34.7	13.2
Arthritis	236	38.4	13.0
Cancer	319	24.8	14.9
Fibromyalgia	303	40.0	12.2
HIV	232	30.0	13.9
Kidney disease	354	31.1	15.3
Multiple sclerosis	255	38.4	11.4
Osteoporosis	171	32.7	12.3
Spinal cord injury	281	25.5	14.1
Alzheimer disease	191	29.1	14.8
Intellectual disability	241	32.1	13.2
Parkinson disease	221	19.8	18.0
Mean (SD)	259.3 (49.1)	31.1 (7.7)	13.7 (1.7)

[a]COPD: chronic obstructive pulmonary disease.

Qualitative Analysis

A secondary aim of this study was to identify potential patterns, consistencies, and gaps in AI-generated exercise recommendation text outputs. Major observations derived from qualitative evaluation of AI-generated exercise recommendations can be found in Multimedia Appendix 2. Briefly, several recurring themes emerged among the total sample, including liability and safety, preference for aerobic exercise, and inconsistencies in the terminology used for exercise professionals. Importantly, AI-generated output showed potential bias and discrimination against certain age-based populations and individuals with disabilities. The implications of these findings are discussed in detail below.

Discussion

Principal Findings

This study sought to explore the suitability of AI-generated exercise recommendations using a popular generative AI platform, ChatGPT. Given the recent launch and popularity of ChatGPT and other similar generative AI platforms, the overall goal was to formally appraise the suitability and readability of AI-generated output likely to be seen by patients and inform exercise and health care professionals and other stakeholders on the potential benefits and limitations of using AI to leverage for patient education. The major findings were that AI-generated output (1) presented 41.2% (107/260) of the content provided in a gold-standard exercise recommendation indicating poor comprehensiveness; (2) of the content provided, chat output was 90.7% (146/161) accurate with most discordance related

to the need for exercise preparticipation health screening; and (3) had college-level readability.

The results of this study are consistent with a recently published research letter that evaluated the appropriateness of CVD prevention recommendations from ChatGPT [40]. Sarraju et al [40] developed 25 questions on fundamental heart disease concepts, posed them to the AI interface, and subjectively graded responses as "appropriate" or "inappropriate." AI-generated responses were deemed to be 84% appropriate with noted misinformation provided for questions surrounding ideal exercise volume and type for health and heart disease prevention. This study expands upon these findings by focusing on ExRx, testing additional metrics (ie, comprehensiveness and readability) using an objective, formal coding system based on a ground truth source, and in an expanded list of clinical populations.

Real-World Implications of These Findings

Our findings suggest that while AI-generated exercise recommendations are generally accurate (146/161, 90.7%), they may lack comprehensiveness in certain critical components of ExRx such as target frequency, intensity, time, and type of exercise, which could potentially hinder ease of implementation or their effectiveness. The most common (ie, 8/15, 53%) source of misinformation was the recommendation to seek medical clearance prior to engaging in any exercise. Potential downstream implications are undue patient concern and triggering an unnecessary number of adults for medical evaluation, both posing as potential barriers to exercise adoption [41,42].

The ACSM preparticipation screening guidelines emphasize the public health message that exercise is important for all individuals and that the preparticipation health screening should not be a deterrent to exercise participation [41]. The preparticipation screening algorithm considers current physical activity levels, desired exercise intensity, and the presence of known or underlying CVD, metabolic, and renal disease. Following this algorithm, lesser than 3% of the general population would be referred before beginning vigorous exercise, and approximately 54% would be referred before beginning any exercise [42]. Interestingly, exercise professionals are well-equipped to facilitate preparticipation screening, yet AI-generated output disproportionately emphasized medical clearance by a health care provider or doctor prior to working with an exercise professional. In reference to exercise professionals, ChatGPT used varying and incorrect terminology such as "licensed exercise physiologist" that does not reflect current-state credentialing for exercise professionals working with clinical populations (ie, ACSM Certified Clinical Exercise Physiologist [43]). These findings corroborate with existing challenges in the public health's understanding of the role of exercise professionals, levels of qualification, and respective scope of practice [44].

As AI-based technologies continue to evolve, striking the right balance between medical precision and risk mitigation remains a crucial consideration [45]. The question of how definitive an AI-based model should be when delivering medical education is multifaceted. On the one hand, the inclination of the AI-based model toward vague or general recommendations can be seen as a responsible stance to mitigate risks. On the other hand, there is merit in AI-based models providing clear, specific, and contextual guidance that reinforces evidence-based recommendations. This approach ensures that end users receive accurate and tailored advice, which is important in the context of medical education. This tension highlights the need for continued dialogue on how AI can enhance health care while ensuring that recommendations align with the highest standards of accuracy and patient safety. These discussions will be instrumental in shaping the future of AI-augmented health care.

AI-Generated Output Least Accurate for Populations With Hypertension

Interestingly, the hypertension exercise recommendations scored the poorest (ie, highest discordance) with 57% (4/7) accuracy and misinformation surrounding the need for medical clearance and the recommended intensity and type of exercise (Table 2). For example, AI-generated output recommended avoiding high-intensity exercise "such as sprinting or heavy lifting"; however, the ACSM does not contraindicate vigorous-intensity exercise considering comorbidities and assuming adequate progression and proper technique [1]. Additionally, AI-generated output recommended a target exercise goal of "30 minutes of moderate-intensity aerobic exercise most days of the week." Notably, the ACSM guidelines reinforce that emphasis is no longer placed on aerobic exercise alone but rather recommend aerobic and resistance exercise alone or combined (ie, concurrent exercise) on most, preferably all, days of the week to total 90-150 minutes per week or more of multimodal, moderate-intensity exercise [39]. Reasons for this discordance are likely because the ChatGPT model relies on training data preceding 2021 and may not capture real-time research advancements. Nevertheless, these findings are important because hypertension is the most common, costly, and modifiable CVD risk factor with strong evidence-based and guideline-driven recommendations, whereby support of exercise is a critical component of first-line treatment for elevated blood pressure [7,46-48].

Social Determinants of Health Considerations

Not surprisingly, our evaluation of this AI-based technology identified social determinants of health considerations regarding educational obtainment for its users. Average readability of the AI-generated output was found to be very high, at the college level, which poses significant challenges for the majority of patients, as The National Institutes of Health, American Medical Association, and American Heart Association all recommend that patient education materials be written at or below a sixth-grade reading level [49] based on national educational obtainment trends. Poor readability of patient materials can exacerbate disparities in access to care for those with limited health literacy, and those individuals may experience more barriers to understand and apply the information provided [29,30]. These findings highlight the need for ongoing evaluation and refinement of AI-generated educational output to prevent inappropriate recommendations that do not improve disparities in clinical outcomes. AI-based models, such as ChatGPT, and their output are vulnerable to both poor data

quality and noninclusive design. Notably, AI-generated output used different tenses and pronouns depending on the demographic group being addressed, which potentially perpetuates digital discrimination including stereotypes and biases (Multimedia Appendix 2). For instance, most AI-generated exercise recommendations were provided in the second-person tense; however, recommendations for individuals with intellectual disabilities, older adults, and children and adolescents were written in the third-person tense with the AI-based model, assuming these populations were not the primary end users. Additionally, most exercise examples provided by the chatbot were activities favoring ambulating individuals (eg, walking and running) potentially limiting education for, and perpetuating bias against, individuals with disabilities. Generative AI can contribute to bias or discrimination in several ways, beginning with the use of biased data to train AI-based models that learn and perpetuate biases in its output [50]. Additionally, AI-based models may be designed with certain features that result in biased or discriminatory outputs, such as using certain variables that are correlated with gender or race [50]. Put in practice, AI-based models can further extend societal biases and stereotypes by relying on existing patterns and trends in the data that reinforce gender or racial stereotypes [50]. These findings highlight the need for caution in using generative AI for health education and the importance of careful consideration of potential biases and discriminatory language.

To summarize, this study demonstrates that AI-generated exercise recommendations hold some promise in accurately providing exercise information but are not without issues (ie, gaps in critical information, biases, and discrimination) that could lead to potentially harmful consequences. The art of ExRx involves considering individual factors and nuances that may not be fully captured by technology [1]. Factors such as medical history, medications, personal preferences, health and physical literacy, and physical limitations are just a few examples of the complexities involved in creating an individualized exercise plan [1]. It is important to note that AI-generated output often lacks references to primary sources or literature, underscoring the need for health care provider oversight in interpreting and verifying the validity of the information presented. In this study, the reference sources provided were 100% accurate (2 of 2); however, “hallucinations” of fabricated or inaccurate references are quite common and are a growing concern for AI-generated medical content [51].

Limitations

There are limitations to this study. This evaluation was limited to a single generative AI platform, which may not be representative of all LLM programs. Additionally, this study is limited to a specific time period and topic, and the findings may not be generalizable to other topics or time periods. Importantly, this model was evaluated using a single, structured prompt that can potentially lead to overfitting or superficial outputs and compromise generalizability. The lack of exposure to a range of prompts makes it challenging to discern if outcomes truly reflect the model’s capabilities or are specific to the nature of the provided prompt. Given that LLMs can yield varied outcomes based on prompts, this limitation is critical for the interpretation and application of the model’s results across various scenarios. This approach was selected as it most closely recapitulates how a publicly available chatbot would likely be used in a real-world setting by an inexperienced end user (ie, lacking knowledge of prompt methodologies). Indeed, all (N=26) AI-generated exercise recommendations were coherent, contextual, and relevant suggesting that the standardized single prompt was structured to elicit an appropriate response. However, it is likely that additional prompt engineering considerations (ie, specificity, iteration, and roles and goals) will yield incremental capabilities and superior model performance than reported in this study. Future work should consider advanced and diverse prompts to assess the model’s robustness across various scenarios. The results rely on the accuracy of the coders in identifying relevant content and assessing its accuracy. The high level of agreement between raters suggests that the coding scheme was well-defined and easily interpretable; however, there is potential for observer bias due to the raters’ shared mentorship, research training, and educational experiences. It is also worth noting that this study used the Flesch-Kincaid formula to assess readability that has known limitations, such as not accounting for the complexity of ideas and vocabulary and not considering readers' cultural and linguistic backgrounds [36]. This tool was selected due to its objectivity, standardization, and the fact that scores are computationally derived, which lowers the risk of human error, thus rendering it the most appropriate tool to address this research question [36]. Nevertheless, future research may benefit from examining the Flesch-Kincaid formula in conjunction with other measures to gain a more comprehensive understanding of AI-generated output readability.

Despite the noted limitations, this study possesses several strengths. To the best of our knowledge, this study is the first to report on the quality of AI-generated exercise recommendations for individuals across the life span (ie, children and adolescents, healthy adults, and older adults) and for 23 additional clinical populations. A major strength of this study is the use of a formal grading framework with a double-coding system to objectively assess the comprehensiveness and accuracy of the AI-generated exercise recommendations, which extends the literature and increases the reliability and validity of these findings [40]. Adding to its credibility, this grading system was developed and refined by experts in the field of exercise science, including a former associate editor [35], editor, and contributing author [1] of the ACSM GETP (LSP and ALZ). Multiple measures were used to assess the suitability of AI-generated recommendations and its potential for digital discrimination. Recommendations were evaluated by their comprehensiveness, accuracy, and readability, which provided a thorough summarization of the strengths and weaknesses of AI-generated content. The output was compared to well-established evidence-based guidelines (ie, ACSM GETP) as a gold-standard reference, which strengthens the validity of the results. Finally, the standardization of queries in this study minimized bias and allowed for an objective evaluation of the AI-generated exercise recommendations. These structured prompts were integral to the research design, shaping the language model's responses and enabling the systematic evaluation of its performance against ACSM GETP as the

ground truth benchmark. This methodological approach ensures that the outcomes presented in this study are grounded in a consistent and rigorously designed interaction process.

Future Directions

Given the recent development of open-source generative AI technologies, this area is ripe for exploration. However, before proceeding with extensive randomized controlled trials, it is crucial to prioritize the safety and ethical considerations associated with AI-generated medical education. As AI technologies have the potential to impact health disparities, it is essential to carefully evaluate their use to ensure inclusivity and appropriate messaging across demographics [27,52-54]. Further research is needed to develop, test, and implement AI technologies that serve individuals safely, effectively, and ethically without perpetuating bias, discrimination, or causing harm. This includes exploring ways to mitigate potential biases and discriminatory outcomes. Outside of the research setting, health care and exercise professionals can play a crucial role in improving AI-based models through prompting and by giving corrective feedback to retrain biases and inaccuracies in AI-generated responses. By enriching ChatGPT with user-specific data including exercise components, literacy level, physical limitations, and other activity considerations, there are opportunities to improve the personalization of recommendations and lessen digital discrimination. Through this stewardship, continuous refinement will likely improve the performance, usability, and appropriateness of the model, translating to superior patient outcomes, which is the goal of provider-enablement and patient-facing tools. As LLMs continue to evolve, it will become increasingly important for researchers to continuously assess improvements with response variations over time. Importantly, future work should explore the incremental value of advanced and diverse prompting considerations. Examples of prompting considerations include the provision of roles and goals (eg, “You are a Clinical Exercise Physiologist and your goal is to design a safe and effective exercise prescription to lower blood pressure”), engaging in multiple or chain prompting and specifically prompting for content commonly missing from output as identified in this study.

To ensure the responsible and safe deployment of AI technologies in health care, conducting thorough implementation studies is a logical next step. These studies should focus on measuring various factors, including acceptability, adoption, appropriateness, costs, feasibility, fidelity, penetration, and sustainability. By thoroughly investigating these implementation aspects, we can ensure that the technology is well-integrated and does not pose any harm to patients or health care systems. Following the completion of the implementation studies, it is important to assess the impact of AI-generated models on service outcomes. This includes evaluating health care quality factors such as safety, timeliness, efficiency, effectiveness, equity, and patient-centeredness [55]. Understanding how AI technologies influence these service outcomes will provide valuable insights into their overall impact on health care delivery. Additionally, measuring patient-centered and end-user outcomes is essential to evaluate the effectiveness of AI technologies in improving patient experiences and outcomes. Randomized controlled trials designed to test ChatGPT as an intervention to augment behavior change and associated health outcomes would be of great public health interest. These trials should prioritize patient-centered outcomes, including satisfaction, usability, experience, and patient activation [56]. By assessing these outcomes, we can determine the effectiveness of AI technologies in empowering patients and fostering meaningful engagement with health care providers.

Conclusions

To conclude, this study found that AI-generated exercise recommendations have moderate comprehensiveness and high accuracy when compared to a gold-standard reference source. However, there are notable gaps in content surrounding critical components of ExRx and potentially biased and discriminatory outputs. Additionally, the readability level of the recommendations may be too high for some patients, and the lack of references in AI-generated content may be a significant limitation for use. Health care providers and patients may wish to remain cautious in relying solely on AI-generated exercise recommendations and should limit their use in combination with clinical expertise and oversight.

Acknowledgments

This study was supported by the University of Connecticut, CVS Health Corporation, and Hartford Hospital.

Authors' Contributions

ALZ contributed to the study conceptualization, project management, study design, data curation and coding, statistical analysis, interpretation of the data, visual presentation of the data, and paper preparation and submission. RB contributed to the study design, data coding, interpretation of the data, and copyediting of the paper. KJTC contributed to the interpretation of the data, business leadership, and copyediting of the paper. LSP contributed to the study design, project oversight, interpretation of the data, and revising and copyediting of the paper. All authors contributed to the writing of the paper, reviewed and approved the final version of the paper, and agreed with the order of presentation of the authors.

Conflicts of Interest

ALZ and KJTC are both employed and hold stock with CVS Health Corporation. This study is an objective evaluation to better understand ChatGPT and its outputs. To the best of our knowledge, CVS Health does not currently use or endorse the use of ChatGPT for lifestyle recommendations. LSP is the sole proprietor and founder of P3-EX, LLC, which could potentially benefit

from the tool used in this research. The results of this study do not constitute endorsement by the American College of Sports Medicine.

Multimedia Appendix 1

Output from artificial intelligence–generated exercise recommendations for clinical populations (N=26).
[PDF File (Adobe PDF File), 243 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Summary of major themes derived from artificial intelligence–generated exercise recommendations.
[PDF File (Adobe PDF File), 116 KB-Multimedia Appendix 2]

References

1. Liguori G. ACSM's Guidelines for Exercise Testing and Prescription. 11th Edition. Philadelphia, PA. Wolters Kluwer; 2021.
2. Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, et al. The physical activity guidelines for Americans. JAMA. 2018;320(19):2020-2028. [FREE Full text] [doi: 10.1001/jama.2018.14854] [Medline: 30418471]
3. Joseph JJ, Deedwania P, Acharya T, Aguilar D, Bhatt DL, Chyun DA, et al. Comprehensive management of cardiovascular risk factors for adults with type 2 diabetes: a scientific statement from the American Heart Association. Circulation. Mar 2022;145(9):e722-e759. [FREE Full text] [doi: 10.1161/CIR.0000000000001040] [Medline: 35000404]
4. Lloyd-Jones DM, Allen NB, Anderson CAM, Black T, Brewer LC, Foraker RE, et al. Life's essential 8: updating and enhancing the American Heart Association's construct of cardiovascular health: a presidential advisory from the American Heart Association. Circulation. Aug 02, 2022;146(5):e18-e43. [FREE Full text] [doi: 10.1161/CIR.0000000000001078] [Medline: 35766027]
5. Pedersen BK, Saltin B. Exercise as medicine—evidence for prescribing exercise as therapy in 26 different chronic diseases. Scand J Med Sci Sports. Dec 2015;25(Suppl 3):1-72. [FREE Full text] [doi: 10.1111/sms.12581] [Medline: 26606383]
6. Pescatello LS, Buchner DM, Jakicic JM, Powell KE, Kraus WE, Bloodgood B, et al. Physical activity to prevent and treat hypertension: a systematic review. Med Sci Sports Exerc. Jun 2019;51(6):1314-1323. [FREE Full text] [doi: 10.1249/MSS.0000000000001943] [Medline: 31095088]
7. Barone Gibbs B, Hivert MF, Jerome GJ, Kraus WE, Rosenkranz SK, Schorr EN, et al. Physical activity as a critical component of first-line treatment for elevated blood pressure or cholesterol: who, what, and how?: A scientific statement from the American Heart Association. Hypertension. Aug 2021;78(2):e26-e37. [FREE Full text] [doi: 10.1161/HYP.0000000000000196] [Medline: 34074137]
8. Exercise is medicine. ACSM's Rx for health. American College of Sports Medicine. 2021. URL: https://www.exerciseismedicine.org/ [accessed 2023-05-01]
9. O'Brien MW, Shields CA, Oh PI, Fowles JR. Health care provider confidence and exercise prescription practices of Exercise is Medicine Canada workshop attendees. Appl Physiol Nutr Metab. Apr 2017;42(4):384-390. [FREE Full text] [doi: 10.1139/apnm-2016-0413] [Medline: 28177736]
10. Fowles JR, O'Brien MW, Solmundson K, Oh PI, Shields CA. Exercise is Medicine Canada physical activity counselling and exercise prescription training improves counselling, prescription, and referral practices among physicians across Canada. Appl Physiol Nutr Metab. May 2018;43(5):535-539. [FREE Full text] [doi: 10.1139/apnm-2017-0763] [Medline: 29316409]
11. Omura JD, Bellissimo MP, Watson KB, Loustalot F, Fulton JE, Carlson SA. Primary care providers' physical activity counseling and referral practices and barriers for cardiovascular disease prevention. Prev Med. Mar 2018;108:115-122. [FREE Full text] [doi: 10.1016/j.ypmed.2017.12.030] [Medline: 29288783]
12. Choudhury A, Asan O, Alelyani T. Exploring the role of the internet, care quality and communication in shaping mental health: analysis of the Health Information National Trends Survey. IEEE J Biomed Health Inform. Jan 2022;26(1):468-477. [doi: 10.1109/JBHI.2021.3087083] [Medline: 34097623]
13. Finney Rutten LJ, Blake KD, Greenberg-Worisek AJ, Allen SV, Moser RP, Hesse BW. Online health information seeking among US adults: measuring progress toward a Healthy People 2020 objective. Public Health Rep. 2019;134(6):617-625. [FREE Full text] [doi: 10.1177/0033354919874074] [Medline: 31513756]
14. Swoboda CM, Van Hulle JM, McAlearney AS, Huerta TR. Odds of talking to healthcare providers as the initial source of healthcare information: updated cross-sectional results from the Health Information National Trends Survey (HINTS). BMC Fam Pract. Aug 29, 2018;19(1):146. [FREE Full text] [doi: 10.1186/s12875-018-0805-7] [Medline: 30157770]
15. Bernard R, Bowsher G, Sullivan R, Gibson-Fall F. Disinformation and epidemics: anticipating the next phase of biowarfare. Health Secur. 2021;19(1):3-12. [FREE Full text] [doi: 10.1089/hs.2020.0038] [Medline: 33090030]
16. Liu T, Xiao X. A framework of AI-based approaches to improving eHealth literacy and combating infodemic. Front Public Health. 2021;9:755808. [FREE Full text] [doi: 10.3389/fpubh.2021.755808] [Medline: 34917575]

17. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. Mar 30, 2023;388(13):1233-1239. [FREE Full text] [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
18. Sezgin E, Sirrianni J, Linwood SL. Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of Generative Pretrained Transformer 3 (GPT-3) as a service model. JMIR Med Inform. Feb 10, 2022;10(2):e32875. [FREE Full text] [doi: 10.2196/32875] [Medline: 35142635]
19. No authors listed. Will ChatGPT transform healthcare? Nat Med. Mar 2023;29(3):505-506. [FREE Full text] [doi: 10.1038/s41591-023-02289-5] [Medline: 36918736]
20. Hu K. ChatGPT sets record for fastest-growing user base-analyst note. Reuters. 2023. URL: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ [accessed 2023-02-02]
21. Chow JCL, Sanders L, Li K. Impact of ChatGPT on medical chatbots as a disruptive technology. Front Artif Intell. 2023;6:1166014. [FREE Full text] [doi: 10.3389/frai.2023.1166014] [Medline: 37091303]
22. Li R, Kumar A, Chen JH. How chatbots and large language model artificial intelligence systems will reshape modern medicine: fountain of creativity or Pandora's box? JAMA Intern Med. Jun 01, 2023;183(6):596-597. [doi: 10.1001/jamainternmed.2023.1835] [Medline: 37115531]
23. Haupt CE, Marks M. AI-generated medical advice—GPT and beyond. JAMA. Apr 25, 2023;329(16):1349-1350. [doi: 10.1001/jama.2023.5321] [Medline: 36972070]
24. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Med Educ. Mar 06, 2023;9:e46885. [FREE Full text] [doi: 10.2196/46885] [Medline: 36863937]
25. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare (Basel). 2023;11(6):887. [FREE Full text] [doi: 10.3390/healthcare11060887] [Medline: 36981544]
26. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. Feb 08, 2023;9:e45312. [FREE Full text] [doi: 10.2196/45312] [Medline: 36753318]
27. Thomas Craig KJ, Morgan LC, Chen CH, Michie S, Fusco N, Snowdon JL, et al. Systematic review of context-aware digital behavior change interventions to improve health. Transl Behav Med. May 25, 2021;11(5):1037-1048. [FREE Full text] [doi: 10.1093/tbm/ibaa099] [Medline: 33085767]
28. Brewer LC, Fortuna KL, Jones C, Walker R, Hayes SN, Patten CA, et al. Back to the future: achieving health equity through health informatics and digital health. JMIR Mhealth Uhealth. Jan 14, 2020;8(1):e14512. [FREE Full text] [doi: 10.2196/14512] [Medline: 31934874]
29. Nutbeam D, Lloyd JE. Understanding and responding to health literacy as a social determinant of health. Annu Rev Public Health. Apr 01, 2021;42:159-173. [FREE Full text] [doi: 10.1146/annurev-publhealth-090419-102529] [Medline: 33035427]
30. Stormacq C, Van den Broucke S, Wosinski J. Does health literacy mediate the relationship between socioeconomic status and health disparities? Integrative review. Health Promot Int. Oct 01, 2019;34(5):e1-e17. [doi: 10.1093/heapro/day062] [Medline: 30107564]
31. ChatGPT Feb 13 version. Open AI. 2023. URL: https://chat.openai.com/chat [accessed 2023-12-21]
32. Pronouncements & scientific communications. American College of Sports Medicine. 2023. URL: https://www.acsm.org/education-resources/pronouncements-scientific-communications [accessed 2023-12-21]
33. Campbell DJ, Estephan LE, Mastrolonardo EV, Amin DR, Huntley CT, Boon MS. Evaluating ChatGPT responses on obstructive sleep apnea for patient education. J Clin Sleep Med. Dec 01, 2023;19(12):1989-1995. [doi: 10.5664/jcsm.10728] [Medline: 37485676]
34. Tabone W, de Winter J. Using ChatGPT for human-computer interaction research: a primer. R Soc Open Sci. Sep 2023;10(9):231053. [FREE Full text] [doi: 10.1098/rsos.231053] [Medline: 37711151]
35. American College of Sports Medicine. ACSM's Guidelines for Exercise Testing and Prescription. 9th Edition. Philadelphia, PA. Wolters Kluwer/Lippincott Williams & Wilkins; 2014.
36. Wang LW, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm. 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
37. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis: striving to meet the trustworthiness criteria. Int J Qual Methods. Oct 02, 2017;16(1):160940691773384. [FREE Full text] [doi: 10.1177/1609406917733847]
38. Ceccarelli G, Pinacchio C, Santinelli L, Adami PE, Borrazzo C, Cavallari EN, et al. Physical activity and HIV: effects on fitness status, metabolism, inflammation and immune-activation. AIDS Behav. Apr 2020;24(4):1042-1050. [doi: 10.1007/s10461-019-02510-y] [Medline: 31016505]
39. Alves AJ, Wu Y, Lopes S, Ribeiro F, Pescatello LS. Exercise to treat hypertension: late breaking news on exercise prescriptions that FITT. Curr Sports Med Rep. Aug 01, 2022;21(8):280-288. [FREE Full text] [doi: 10.1249/JSR.0000000000000983] [Medline: 35946847]

40. Sarraju A, Bruemmer D, Van Iterson E, Cho L, Rodriguez F, Laffin L. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. JAMA. Mar 14, 2023;329(10):842-844. [FREE Full text] [doi: 10.1001/jama.2023.1044] [Medline: 36735264]
41. Riebe D, Franklin BA, Thompson PD, Garber CE, Whitfield GP, Magal M, et al. Updating ACSM's recommendations for exercise preparticipation health screening. Med Sci Sports Exerc. Nov 2015;47(11):2473-2479. [FREE Full text] [doi: 10.1249/MSS.0000000000000664] [Medline: 26473759]
42. Whitfield GP, Riebe D, Magal M, Liguori G. Applying the ACSM preparticipation screening algorithm to U.S. adults: National Health and Nutrition Examination Survey 2001-2004. Med Sci Sports Exerc. Oct 2017;49(10):2056-2063. [FREE Full text] [doi: 10.1249/MSS.0000000000001331] [Medline: 28557860]
43. Which certification is right for you? American College of Sports Medicine. 2023. URL: https://www.acsm.org/certification/get-certified [accessed 2023-05-18]
44. Gallo PM. The United States Registry for Exercise Professionals: how it works and ways it can advance the fitness profession. ACSM's Health Fitness J. 2023;27(2):51-53. [doi: 10.1249/fit.0000000000000843]
45. Mohammad B, Supti T, Alzubaidi M, Shah H, Alam T, Shah Z, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform. Jun 29, 2023;305:644-647. [doi: 10.3233/SHTI230580] [Medline: 37387114]
46. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. Sep 10, 2019;140(11):e596-e646. [FREE Full text] [doi: 10.1161/CIR.0000000000000678] [Medline: 30879355]
47. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart Disease and Stroke Statistics-2022 update: a report from the American Heart Association. Circulation. Feb 22, 2022;145(8):e153-e639. [FREE Full text] [doi: 10.1161/CIR.0000000000001052] [Medline: 35078371]
48. Hanssen H, Boardman H, Deiseroth A, Moholdt T, Simonenko M, Kränkel N, et al. Personalized exercise prescription in the prevention and treatment of arterial hypertension: a Consensus Document from the European Association of Preventive Cardiology (EAPC) and the ESC Council on Hypertension. Eur J Prev Cardiol. Feb 19, 2022;29(1):205-215. [FREE Full text] [doi: 10.1093/eurjpc/zwaa141] [Medline: 33758927]
49. Siddiqui E, Shah AM, Sambol J, Waller AH. Readability assessment of online patient education materials on atrial fibrillation. Cureus. Sep 11, 2020;12(9):e10397. [FREE Full text] [doi: 10.7759/cureus.10397] [Medline: 33062517]
50. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI Soc. 2023;38(2):549-563. [FREE Full text] [doi: 10.1007/s00146-022-01455-6] [Medline: 35615443]
51. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. Cureus. May 2023;15(5):e39238. [FREE Full text] [doi: 10.7759/cureus.39238] [Medline: 37337480]
52. Garvey KV, Craig KJT, Russell RG, Novak L, Moore D, Preininger AM, et al. The potential and the imperative: the gap in AI-related clinical competencies and the need to close it. Med Sci Educ. Dec 2021;31(6):2055-2060. [FREE Full text] [doi: 10.1007/s40670-021-01377-w] [Medline: 34956712]
53. Garvey KV, Thomas Craig KJ, Russell R, Novak LL, Moore D, Miller BM. Considering clinician competencies for the implementation of artificial intelligence-based tools in health care: findings from a scoping review. JMIR Med Inform. Nov 16, 2022;10(11):e37478. [FREE Full text] [doi: 10.2196/37478] [Medline: 36318697]
54. Novak LL, Russell RG, Garvey K, Patel M, Thomas Craig KJ, Snowdon J, et al. Clinical use of artificial intelligence requires AI-capable organizations. JAMIA Open. Jul 2023;6(2):ooad028. [FREE Full text] [doi: 10.1093/jamiaopen/ooad028] [Medline: 37152469]
55. Thomas Craig KJ, McKillop MM, Huang HT, George J, Punwani ES, Rhee KB. U.S. hospital performance methodologies: a scoping review to identify opportunities for crossing the quality chasm. BMC Health Serv Res. Jul 10, 2020;20(1):640. [FREE Full text] [doi: 10.1186/s12913-020-05503-z] [Medline: 32650759]
56. Bruce C, Harrison P, Giammattei C, Desai SN, Sol JR, Jones S, et al. Evaluating patient-centered mobile health technologies: definitions, methodologies, and outcomes. JMIR Mhealth Uhealth. Nov 11, 2020;8(11):e17577. [FREE Full text] [doi: 10.2196/17577] [Medline: 33174846]

Abbreviations

ACSM: American College of Sports Medicine
AI: artificial intelligence
CVD: cardiovascular disease
ExRx: exercise prescription
FITT: frequency, intensity, time, and type
GETP: Guidelines for Exercise Testing and Prescription

GPT: generative pretrained transformer
LLM: large language model

Edited by G Eysenbach, K Venkatesh, MN Kamel Boulos; submitted 27.07.23; peer-reviewed by A Sarraju, M Mahling; comments to author 16.09.23; revised version received 05.10.23; accepted 11.12.23; published 11.01.24

Please cite as:
Zaleski AL, Berkowsky R, Craig KJT, Pescatello LS
Comprehensiveness, Accuracy, and Readability of Exercise Recommendations Provided by an AI-Based Chatbot: Mixed Methods Study
JMIR Med Educ 2024;10:e51308
URL: https://mededu.jmir.org/2024/1/e51308
doi: 10.2196/51308
PMID: 38206661